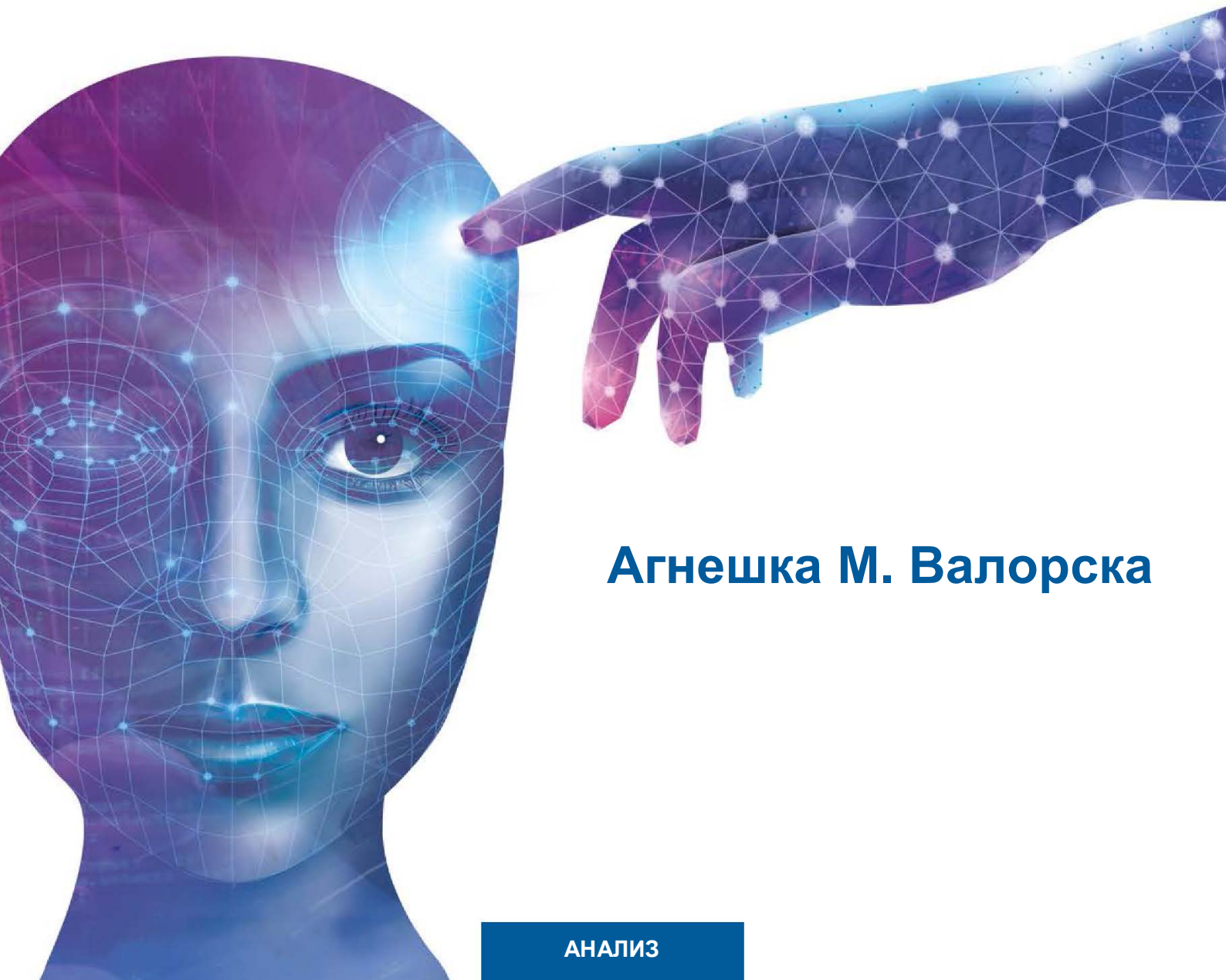
FRIEDRICH NAUMANN FOUNDATION For Freedom.

ДИПФЕЙКИ
И ДЕЗИНФОРМАЦИЯ



Агнешка М. Валорска

АНАЛИЗ

ВЫХОДНЫЕ ДАННЫЕ

Издатель
Фонд Фридриха Науманна за свободу
Карл-Маркс-Штрассе 2
14482 Потсдам
Германия

/freiheit.org
/FriedrichNaumannStiftungFreiheit
/FNFreiheit

Автор
Агнешка М. Валорска

Редакторы
Международный отдел
Подразделение глобальных тем
Фонд Фридриха Науманна за свободу

Концепция и оформление
TroNa GmbH

Контакт
Телефон: +49 (0)30 2201 2634
Факс: +49 (0)30 6908 8102
Email: service@freiheit.org

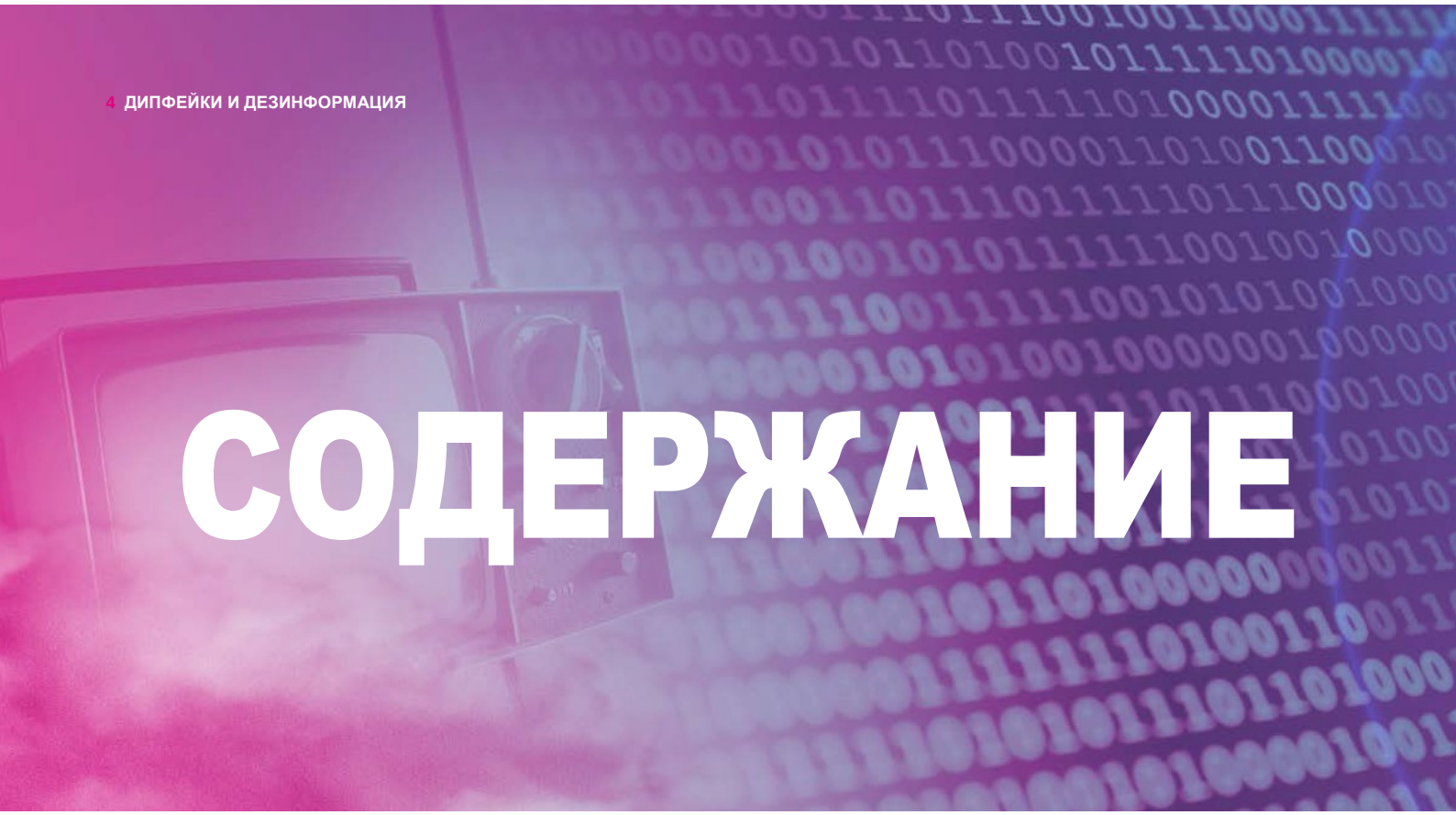
По состоянию на
май 2020 г.

Сведения о фотографиях
Фотомонтажи
© Unsplash.de, © freepik.de, стр. 30 © AdobeStock

Скриншоты
стр. 16 © https://youtu.be/mSaIrz8lM1U
стр. 18 © deepnude.to / Агнешка М. Валорска
стр. 19 © thispersondoesnotexist.com
стр. 19 © linkedin.com
стр. 19 © talktotransformer.com
стр. 25 © gltr.io
стр. 26 © twitter.com

Все остальные фотографии
© Фонд Фридриха Науманна за свободу (Германия)
стр. 31 © Агнешка М. Валорска

Замечания по использованию этой публикации
Данная публикация является информационной услугой Фонда Фридриха Науманна за свободу. Публикация доступна бесплатно и не продается. Она не может использоваться партиями или агитаторами для агитации во время избирательной кампании (при выборах в Бундестаг, региональных и местных выборах или выборах в Европейский парламент).



СОДЕРЖАНИЕ

Оглавление

СВОДНОЕ РЕЗЮМЕ

Применение искусственного интеллекта (ИИ) играет все более важную роль в нашем обществе, но новые возможности этой технологии идут рука об руку с новыми рисками. Одним из таких рисков является ненадлежащее использование технологии для преднамеренного распространения ложной информации. Хотя политически мотивированное распространение дезинформации, конечно, не новое явление, технический прогресс сделал создание и распространение управляемого контента намного проще и эффективнее, чем когда-либо прежде. При использовании алгоритмов ИИ, видео теперь можно быстро и относительно дешево фальсифицировать («*дипфейки*») без наличия каких-либо специальных знаний.

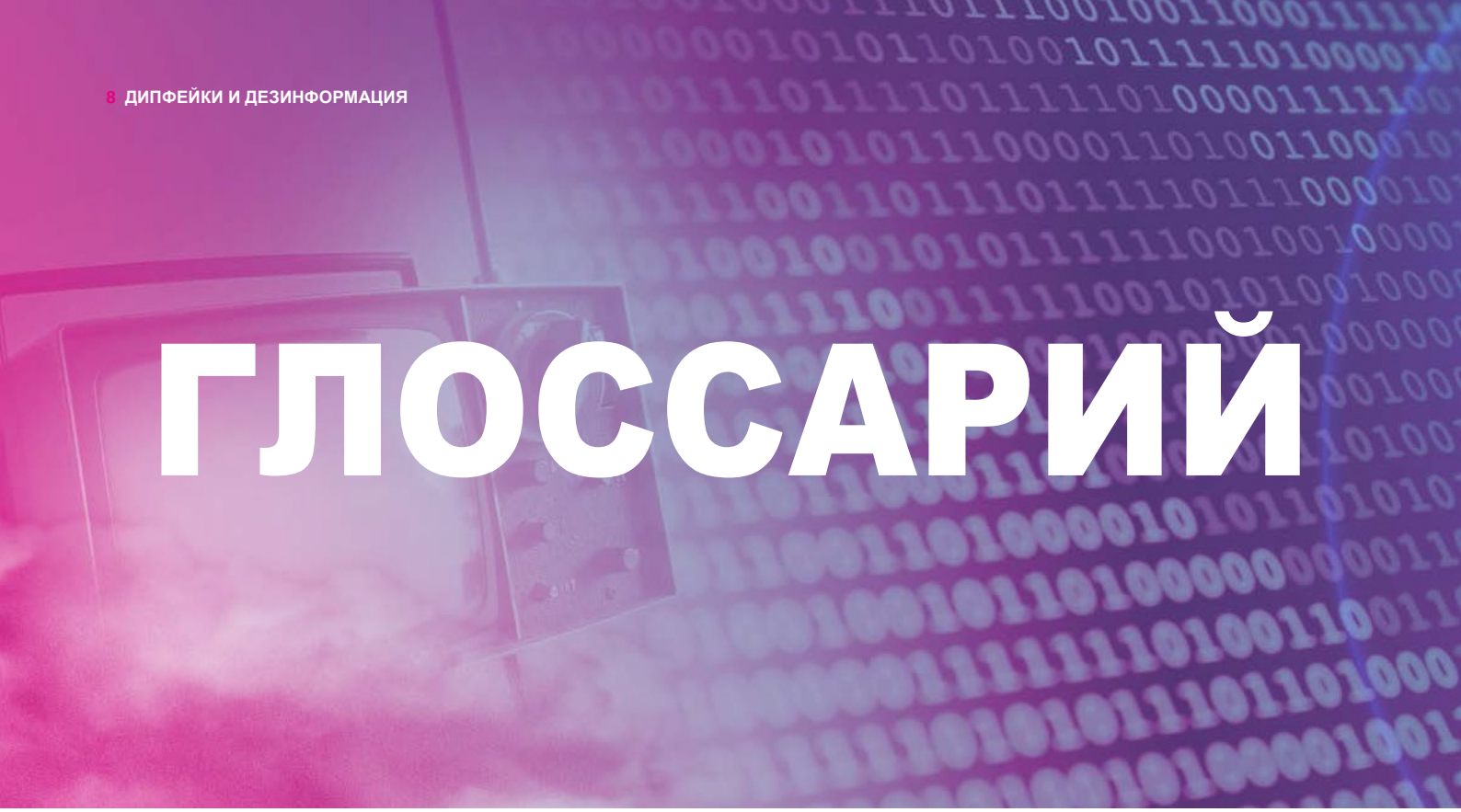
Дискурс на эту тему в первую очередь ориентирован на потенциальное использование дипфейков в избирательных кампаниях, но этот тип видео составляет лишь малую часть всех таких манипуляций: в 96% случаях дипфейки использовались для создания порнографических фильмов с участием известных женщин. Женщины извне публичной сферы также могут оказаться непроизвольными звездами этого вида манипулируемого видео (дипфейковой мести порнографией). Кроме того, такие приложения, как DeepNude, позволяют преобразовывать статические изображения в обманчиво реальные изображения обнаженной натуры. Неудивительно, что эти приложения работают только с изображениями женских тел. Но визуальный контент – это не единственный тип контента, которым можно управлять или который можно создавать алгоритмически. Сгенерированные ИИ голоса уже успешно использовались для совершения мошенничества, что приводило к значительным финансовым убыткам, а GPT-2 может генерировать тексты, содержащие произвольные факты и цитаты.

Как лучше всего решать эти проблемы? Компании и исследовательские институты уже вложили значительные средства в технологические решения для выявления сгенерированных ИИ видеороликов. Преимущество таких вложений обычно недолговечно: разработчики дипфейков реагируют на технологические решения идентификации более изощренными методами – классический пример гонки вооружений. По этой причине платформы, распространяющие манипулируемый контент, должны нести большую ответственность. Facebook и Twitter теперь самостоятельно ввели правила для обработки манипулируемого контента, но эти правила не являются единообразными, и нежелательно оставлять частным компаниям право определять, что влечет за собой «*свобода выражения мнения*».

Федеральное правительство Германии явно не готово к теме «*Применение управляемого ИИ контента в целях дезинформации*», как показал краткий парламентский запрос, представленный парламентской группой FDP в декабре 2019 года. В правительстве нет четко определенной ответственности за этот вопрос и нет конкретного законодательства. До сих пор применялись только «*общие и абстрактные правила*». Ответы федерального правительства не предполагают какой-либо конкретной стратегии или каких-либо намерений инвестировать, чтобы лучше подготовиться к решению этой проблемы. В целом, существующие попытки регулирования на уровне Германии и Европы не кажутся достаточными для решения проблемы дезинформации на основе ИИ. Но это не обязательно должно быть так. В некоторых штатах США уже приняли законы как против неконсенсуальной дипфейковой порнографии, так и против использования этой технологии для влияния на избирателей.

Соответственно, законодатели должны разработать четкие руководящие принципы для цифровых платформ по единообразной обработке дипфейков в частности и дезинформации в целом. Меры могут варьироваться от маркировки манипулируемого контента как такового и ограничения его распространения (исключения его из алгоритмов рекомендаций) до его удаления. Повышение медиаграмотности также должно стать приоритетом для всех граждан, независимо от возраста. Важно повысить осведомленность широкой общественности о существовании дипфейков и развивать способность людей анализировать аудиовизуальный контент, даже если выявлять фейки становится все труднее. В этом отношении стоит отметить подход, принятый северными странами, особенно Финляндией, население которой оказалось наиболее устойчивым к дезинформации.

Тем не менее, есть одна вещь, которую мы не должны делать: поддаваться искушению полностью запретить дипфейки. Как и любая технология, дипфейки открывают множество интересных возможностей – в том числе для образования, кино и сатиры – несмотря на свои риски.



ГЛОССАРИЙ

Общий искусственный интеллект / сильный ИИ

Концепция сильного ИИ или ОИИ относится к компьютерной системе, которая справляется с широким спектром различных задач и, таким образом, достигает уровня интеллекта, подобного человеческому. В настоящее время такого применения ИИ не существует. Например, в настоящее время ни одна система не способна распознавать рак, играть в шахматы и управлять автомобилем, хотя существуют специализированные системы, которые могут выполнять каждую задачу отдельно. Множество исследовательских институтов и компаний в настоящее время работают над сильным ИИ, но нет единого мнения о том, можно ли этого достичь, и если да, то когда.

Big Tech

Термин *«Big Tech»* используется в средствах массовой информации для обозначения группы компаний, занимающих доминирующее положение в ИТ-индустрии. Он часто взаимозаменяется «*GAFA*» или *«большой четверкой»*, подразумевая Google, Apple, Facebook и Amazon (или «GAFAM» , если включать Microsoft). Для китайских крупных технологических компаний используется аббревиатура BATX, подразумевая Baidu, Alibaba, Tencent и Xiaomi.

Чипфейки / Фейки неглубокого залегания

В отличие от дипфейков, фейки неглубокого залегания – это манипуляции с изображениями, аудио или видеороликами, созданные с помощью относительно простых технологий. Примеры включают уменьшение скорости аудиозаписи или отображение контента в измененном контексте.

DARPA

Агентство перспективного планирования научно-исследовательских работ входит в состав Министерства обороны США, и ему поручено исследование и финансирование новаторских военных технологий. В прошлом проекты, финансируемые DARPA, привели к появлению основных технологий, которые также используются в невоенных приложениях, включая Интернет, машинный перевод и беспилотные автомобили.

Дипфейк

Дипфейки (сочетание глубокого обучения и фейка) – это продукт двух алгоритмов ИИ, работающих вместе в так называемой генеративно-состязательной сети (GAN). Сети GAN лучше всего описать как способ алгоритмического генерирования новых типов данных из существующих наборов данных. Например, GAN может проанализировать тысячи фотографий Дональда Трампа, а затем сгенерировать новое изображение, которое будет похоже на проанализированные изображения, но не будет точной копией любого из них. Эта технология может применяться к различным типам контента: изображениям, движущимся изображениям, звуку и тексту. Термин «дипфейк» в основном используется для аудио- и видеоконтента.

Глубокое обучение

Глубокое обучение – это подобласть машинного обучения, в которой искусственные нейронные сети обучаются на больших объемах данных. Подобно тому, как люди учатся на собственном опыте, алгоритмы глубокого обучения повторяют задачу, чтобы постепенно улучшать результаты. Это называется глубоким обучением, потому что нейронные сети имеют несколько уровней для обучения. Глубокое обучение позволяет машинам решать сложные задачи даже при использовании неоднородных неструктурированных наборов данных.

Deep Porn

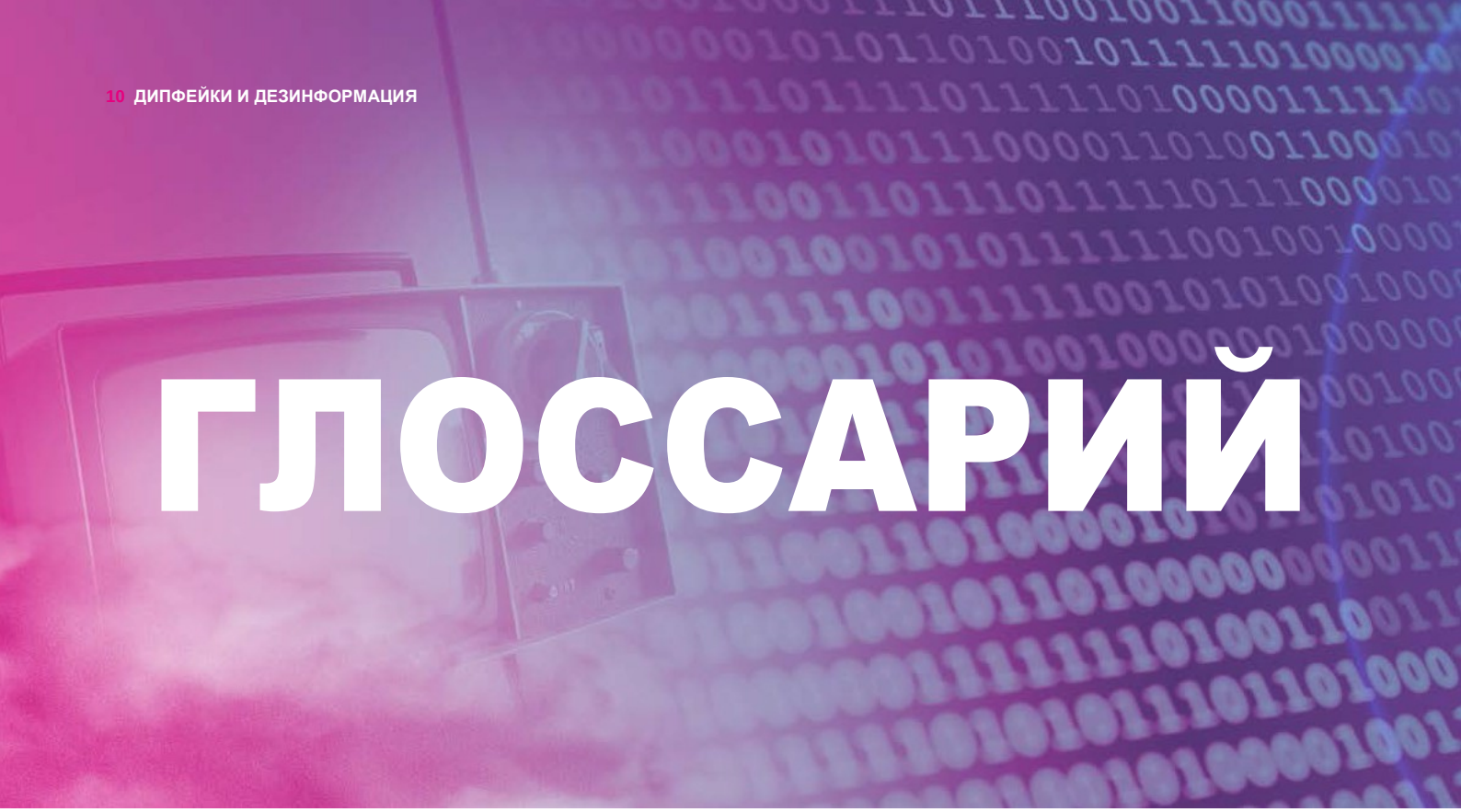
Deep Porn относится к использованию методов глубокого обучения для формирования искусственных порнографических изображений.

Генеративно-состязательные сети

Генеративно-состязательные сети – это алгоритмические архитектуры, основанные на паре из двух нейронных сетей, а именно одной генеративной сети и одной дискриминирующей сети. Эти две сети конкурируют друг против друга (генеративная сеть генерирует данные, а дискриминационная сеть фальсифицирует данные) для создания новых синтетических наборов данных. Процесс повторяется несколько раз, чтобы достичь результатов, которые очень похожи на реальные данные. Сети могут работать с разными типами данных и, следовательно, могут использоваться для создания изображений, а также текста, аудио или видео.

GPT-2

GPT-2 представляет собой каркас на основе искусственной нейронной сети, разработанный исследовательской компанией OpenAI. Он умеет автоматически генерировать тексты на английском языке. Набор данных, на котором основывается GPT-2, содержит около 45 миллионов страниц текста. В отличие от обычных текстовых генераторов, GPT-2 не составляет тексты из готовых текстовых блоков и не ограничивается какой-либо конкретной областью. Он может генерировать новый контент из любого предложения или фрагмента текста.



ГЛОССАРИЙ

IBM Watson

IBM Watson является системой, основанной на машинном обучении, разработанной IBM. Она была разработана с целью создания системы, которая может понимать и отвечать на вопросы, заданные на естественном языке. Watson получила широкое внимание средств массовой информации в 2011 году, когда она победила лучших игроков в телешоу Jeopardy. С тех пор IBM Watson позиционируется как «ИИ для бизнеса», который предлагает ряд облачных и информационных продуктов для различных отраслей: от здравоохранения до кинопроизводства.

Зима искусственного интеллекта

Зима искусственного интеллекта – это период снижения интереса и сокращения финансирования исследований в области искусственного интеллекта. Термин был придуман по аналогии с идеей ядерной зимы. Как технологическая область ИИ с 1950-х годов пережил несколько фаз ажиотажа, за которым последовали разочарование, критика и сокращение финансирования.

Искусственные нейронные сети

Искусственные нейронные сети (ИНС) – это компьютерные системы, в основе которых лежат биологические нейронные сети, обнаруженные в мозге людей и животных. ИНС «*учатся*» выполнять задачи на основе примеров, не запрограммированных какими-либо конкретными правилами.

ИНС могут, например, научиться распознавать изображения, содержащие кошек, анализируя образцы изображений, которые были вручную помечены как «*кошка*» или «*без кошки*», а затем использовать результаты для распознавания кошек на других изображениях.

Машинное обучение

По существу дела, машинное обучение – это метод, который применяет алгоритмы для анализа данных, изучения этих данных и последующего прогнозирования на их основе. Таким образом, вместо ручного программирования программного обеспечения с четко определенными инструкциями для выполнения определенной задачи, программное обеспечение обучается с использованием больших объемов данных и алгоритмов, которые дают ему возможность узнать, как должна выполняться задача.

Микроадресация

Микроадресация представляет собой метод электронного маркетинга, который направлен на сопоставление рекламных кампаний с потребителями или компаниями, которые, скорее всего, будут заинтересованы в продукте или услуге. В зависимости от платформы для этого используются, например, демографические данные целевой аудитории, ее интересы и история просмотров. Основываясь на этих критериях, один и тот же организатор кампании может совершенно по-разному обращаться к разным получателям. Этот маркетинговый инструмент был первоначально разработан для политических кампаний, но сегодня он используется и в коммерческих кампаниях.

Фишинг

Фишинг представляет собой метод кибератаки с помощью электронной почты. Цель состоит в том, чтобы убедить получателя электронной почты в подлинности и актуальности сообщения (например, уведомления из банка), чтобы побудить его нажать ссылку или загрузить вложение. Это позволяет хакерам получить доступ к конфиденциальной информации, такой как пароли.

Порноместь

Порноместь относится к распространению интимных сексуальных изображений или видеороликов без согласия участников. Часто это делается в качестве мести бывшими партнерами после разрыва отношений. Три четверти жертв порномести представляют собой женщины.

Слабый ИИ или специализированный ИИ

Алгоритмы слабого ИИ специализируются на выполнении очень специфических задач, таких как распознавание лиц, понимание языка или игра в шахматы. Хотя они, как правило, намного лучше или эффективнее людей справляются с этими задачами, они способны решать только те задачи, для которых они были разработаны. Любое современное применение искусственного интеллекта относится к категории слабого ИИ, даже такие сложные на вид системы, как беспилотные автомобили или языковые помощники.

Социальная инженерия

Под социальной инженерией понимаются любые меры, предполагающие целенаправленное влияние на человека или людей, например, для получения доступа к конфиденциальной информации или для того, чтобы убедить цель произвести платеж. Эта практика также называется «*социальным хакерством*», когда целью является получение доступа к компьютерным системам целевого лица или организации.

Суперинтеллект

Суперинтеллект – это гипотетический сценарий, при котором искусственный интеллект не только превосходит самых умных людей в отдельности, но и заменяет коллективный разум человечества.

Вишинг

Вишинг (голосовой фишинг) представляет собой метод фишинга, использующий телефонные звонки вместо электронной почты. Использование дипфейков для создания голосовых клипов может повысить эффективность этого метода.

1.0 СОСТОЯНИЕ РАЗВИТИЯ

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ЕГО РОЛЬ В ДЕЗИНФОРМАЦИИ

Хотя корни технологии уходят в середину XX века, искусственному интеллекту долгое время уделялось мало внимания. Долгая зима искусственного интеллекта начала ослабевать только в начале 2010-х годов. В 2011 году компьютерная система IBM Watson победила лучших игроков в телешоу Jeopardy[1], прототипы беспилотных автомобилей Google преодолели более 100.000 миль (160.000 км), а Apple представила своего «*умного личного помощника*» Siri. С тех пор общественный интерес к искусственному интеллекту, особенно к связанным с ним рискам, неуклонно растет. Дискурс о суперинтеллекте, вызванный одноименной книгой Ника Бострома, опубликованной в 2014 году, привлек еще больше внимания. С тех пор известные личности неоднократно предупреждали об ИИ, иногда приобретая тревожную интонацию.

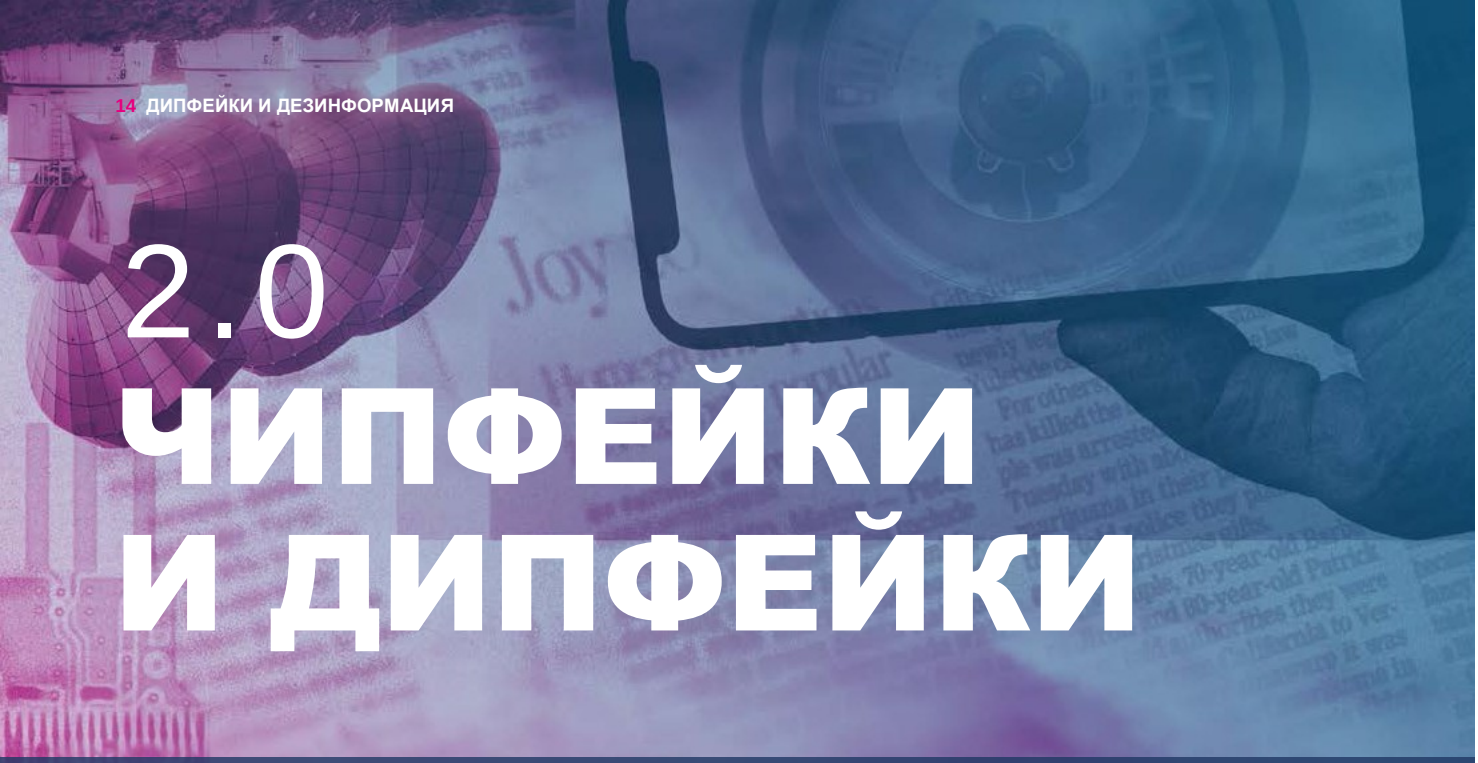
Часто цитируют *Стивена Хокинга («*Разработка полностью искусственного интеллекта может означать конец человеческой расы*») и Илона Маска («*Искусственный интеллект является фундаментальным риском для существования человеческой цивилизации*»).* В то время как супер-интеллект и так называемый «*сильный ИИ*» (ОИИ, общий искусственный интеллект) все еще в далеком будущем, «*слабый ИИ*» и его, возможно, не такие уж слабые алгоритмы уже играют постоянно растущую роль в бизнесе, обществе и политике. По мнению автора, воздействие на здоровье, энергию, безопасность, мобильность и многие другие области будет в значительной степени положительным. Однако мы сможем насладиться положительными аспектами этих разработок, только если мы осознаем риски, связанные с этой технологией, и будем успешно им противодействовать.

1 Jeopardy - это телевикторина, в которой участники получают общие подсказки, представленные в качестве ответа, и должны сформулировать свой ответ в форме вопроса. На протяжении многих лет немецкие адаптации включали «*Riskant on RTL*» и «*Der Große Preis on ZDF*».

«Однако мы сможем насладиться положительными аспектами этих разработок, только если мы осознаем риски, связанные с этой технологией, и будем успешно им противодействовать».

Одним из таких рисков является ненадлежащее использование технологии для преднамеренного распространения ложной информации. Конечно, политически мотивированная дезинформация – явление не новое. Сталин и Мао являются наиболее яркими примерами диктаторов, которые регулярно приказывали редактировать свои фотографии, чтобы старые изображения соответствовали последней «*правде*»: любого, кто впадал в немилость, удаляли с фотографий, новые пополнения в руководстве партии редактировались задним числом; изменялся даже контекст изображений, например, путем изменения фона. Целью манипулирования этими визуальными записями было создание новых фактов, переписывание прошлых событий и самой истории.

Исторически сложилось так, что выполнение этих изменений было утомительным и требовало специальных знаний; сегодня при наличии подходящего приложения для смартфона любой может сделать то же самое без особых усилий. И технология не остановилась на фотографии. Создание фальсифицированного видеоролика, который кажется правдоподобным, по-прежнему требует значительных усилий. Но определенные методы искусственного интеллекта упрощают манипулирование существующими видеороликами. Эти видеоролики стали известны как «*дипфейки*». Они все еще относительно редко встречаются в Интернете, но по мере того, как их использование и распространение увеличивается, они превращаются в растущую проблему для нашего общества. Контент, которым манипулируют, не только очень быстро распространяется на таких платформах, как Facebook или YouTube, но и нацелен на пользователей, которые восприимчивы к нему. Кроме того, дезинформация все больше смещается в сторону служб обмена сообщениями, таких как WhatsApp. Там зашифрованные сообщения распространяются по частным соединениям, это увеличивает доверие к пересылаемой информации, создавая своего рода скрытую виральность. Шифрование частных онлайн-коммуникаций является желательным продуктом, аналогичным секретности написанных писем – оно предотвращает просмотр сообщений третьими лицами. Но шифрование также означает, что любая распространенная информация не может быть проверена на достоверность и соответственно модерирована.



2.0 ЧИПФЕЙКИ И ДИПФЕЙКИ

Технологические возможности подделки текста, изображения, аудио- и видеороликов

За последние два года термин «дипфейк» получил все большее распространение. Но что такое дипфейки, и чем они отличаются от другого манипулируемого контента? Хотя первые научные эксперименты по обработке видео на основе искусственного интеллекта относятся к концу 1990-х годов, широкая публика узнала о технических возможностях только к концу 2017 года.

Это было также тогда, когда была придумана терминология, когда пользователь Reddit по имени «*Deepfakes*» и другие члены сообщества Reddit «*r/deepfakes*» опубликовали созданный ими контент.

Неудивительно, что во многих случаях это использовалось, чтобы сделать порнографические видеоролики, где лица актрис заменяются лицами таких знаменитостей, как Скарлетт Йоханссон или Тейлор Свифт. Более безобидный пример – это снятие сцен из фильма и замена лица каждого актера на лицо Николаса Кейджа.

«Неудивительно, что во многих случаях это использовалось, чтобы сделать порнографические видеоролики, где лица актрис заменяются лицами таких знаменитостей, как Скарлетт Йоханссон или Тейлор Свифт».

ДИПФЕЙКИ РАБОТАЮТ СЛЕДУЮЩИМ ОБРАЗОМ

Дипфейки (сочетание глубокого обучения и фейка) – это продукт двух алгоритмов ИИ, работающих вместе в так называемой генеративно-состязательной сети (GAN). GAN лучше всего охарактеризовать как способ создания новых типов данных из существующих наборов данных алгоритмически.

Например, GAN может проанализировать тысячи фотографий Дональда Трампа, а затем сгенерировать новое изображение, которое будет похоже на проанализированные изображения, но не будет точной копией любого из них. Эта технология может применяться к различным типам контента: изображениям, движущимся изображениям, звуку и тексту. Термин «дипфейк» в основном используется для аудио- и видеоконтента.

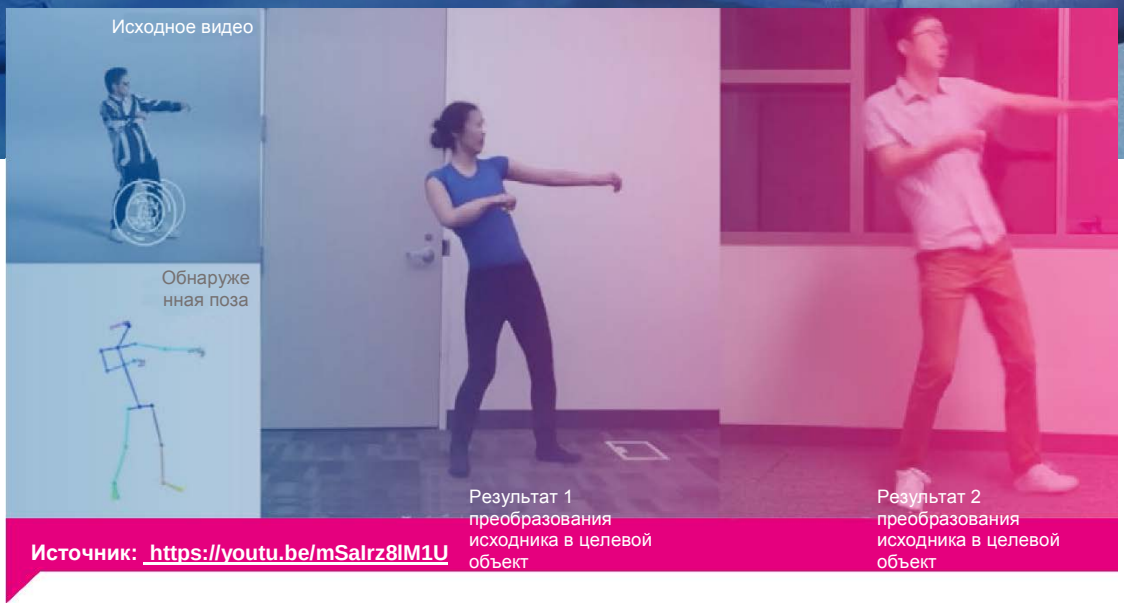
Сегодня для достижения правдоподобных результатов требуется всего несколько сотен изображений или аудиозаписей в качестве данных для обучения. Менее чем за 3 доллара любой человек может заказать фальсифицированный видеоролик с изображением кого-либо по своему выбору при условии, что у него есть как минимум 250 фотографий этого человека, но это вряд ли станет препятствием для человека, использующего Instagram или Facebook. Всего за 10 долларов за 50 слов также могут быть созданы синтетические голосовые записи.

2.1 Дипфейки против чипфейков

Хотя порнографические манипуляции, несомненно, являются наиболее распространенными примерами дипфейков, они не являются основным мотивом для текущей общественной дискуссии. Интересно, что видео, вызвавшее дебаты, ни в коем случае не было дипфейком, а просто чипфейком (иногда также называемым фейком неглубокого залегания): видео спикера Палаты представителей США Нэнси Пелоси, сфальсифицированное очень простыми техническими средствами. Запись была замедлена примерно до 75% от исходной скорости, при этом высота звука была увеличена, чтобы голос по-прежнему звучал естественно. Результаты: У зрителя сложилось правдоподобное впечатление, что Нэнси Пелоси была пьяна.

Этим видеороликом поделились в социальных сетях миллионы раз. Это показывает, как даже самые простые подделки могут исказить реальность и использоваться в политических целях. Тем не менее, исторически было очень сложно фальсифицировать записи, чтобы заставить субъекта выполнять совершенно другие движения или говорить совершенно другие слова, чем в исходном видео. До сих пор.

ПРИМЕРЫ ПРИМЕНЕНИЯ



Источник: https://youtu.be/mSaIrz8lM1U

2.2 Манипулирование паттернами движений

В 2018 году приложение четырех исследователей из Беркли привлекло широкое внимание, оно использовало искусственный интеллект для переноса танцевальных движений от исходного человека (например, профессионального танцора) целевому человеку.[2]

Движения передаются от исходного видео *«фигурке»*. Затем нейронная сеть синтезирует целевое видео в соответствии с «*движениями фигурки*».

В результате получается «*сфальсифицированный*» видеоролик, в котором третье лицо танцует как профессионал. Конечно, этот тип алгоритма можно использовать не только для имитации танцевальных движений, но потенциально для создания любой другой формы движения. Это открывает возможность изображать политических оппонентов в компрометирующих ситуациях: Что, например, может быть разветвлениями видео, в которых политик выполняет нацистское приветствие или просто показывает средний палец?

2 https://arxiv.org/pdf/1808.07371.pdf

ИСКУССТВЕННЫЕ НЕЙРОННЫЕ СЕТИ

Искусственные нейронные сети (ИНС) – это компьютерные системы, в основе которых лежат биологические нейронные сети, обнаруженные в мозге людей и животных.

ИНС «*учатся*» выполнять задачи на основе примеров, не запрограммированных какими-либо конкретными правилами. Они могут, например, научиться распознавать изображения, содержащие кошек, анализируя образцы изображений, которые были вручную помечены как «*кошка*» или «*без кошки*», и использовать результаты для распознавания кошек на других изображениях.

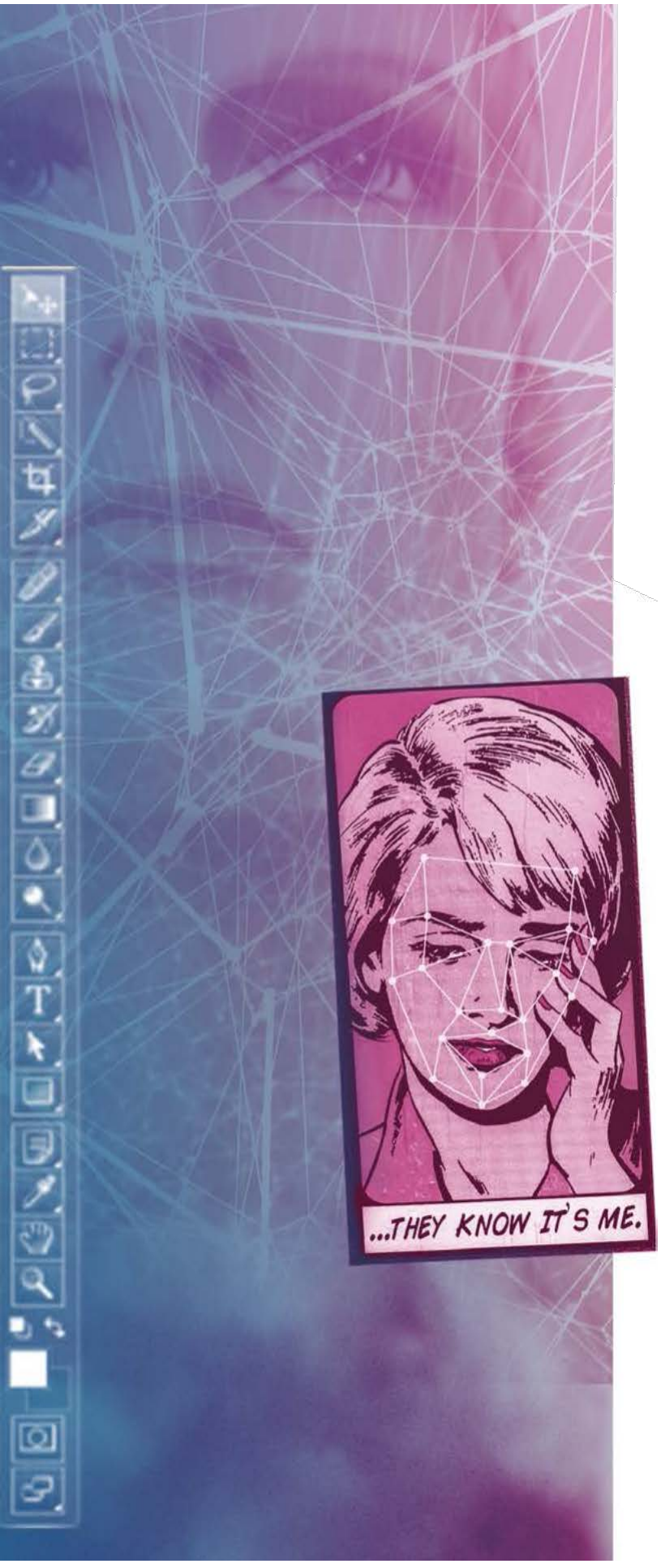
Голос и выражения лица

Подделки могут иметь еще более серьезные последствия, поскольку на них будет казаться, что люди произносят слова, которые никогда не произносились, сопровождаются жестами, мимикой и интонациями голоса, которые кажутся невероятно реалистичными. Была создана серия таких видеороликов, в том числе примеры Барака Обамы и Марка Цукерберга, не для того, чтобы ввести аудиторию в заблуждение, а чтобы продемонстрировать возможности и риски этой технологии. С тех пор был случай, когда дипфейк был создан и распространялся политической партией, Бельгийской социалистической партией – Socialistische Partij Anders (sp.a.).

В мае 2018 года партия разместила в Facebook видеоролик, в котором Трамп высмеивал Бельгию за соблюдение Парижского соглашения по климату.[3]

Несмотря на явно низкое качество и неестественные движения губ, которые должны были вызвать подозрение у любого внимательного зрителя, видео вызвало сотни комментариев, многие из которых выражали возмущение тем, что американский президент посмел вмешаться в политику Бельгии по климату. Создатели этого видео также пытались способствовать пониманию проблемы. Видео было целенаправленной провокацией, чтобы привлечь внимание людей к онлайн-петиции, призывающей бельгийское правительство принять более срочные меры по проблемам климата. Но что, если кто-то создаст видео, в котором Трамп будет говорить не на тему политики Бельгии по климату, например, о своем намерении атаковать Иран?

3 https://www.facebook.com/watch/?v=10155618434657151



Манипуляция изображениями: DeepNude и искусственные лица

Изображение и текстовый контент часто не классифицируются как дипфейки, хотя они могут быть созданы с помощью очень схожей технологии. Для этого есть простая причина: как изображениями, так и текстами можно так легко манипулировать без необходимости сложной технологии, что «*польза*» (или вред, в зависимости от точки зрения) от этого намного меньше, чем при манипуляциях с аудио- и видеоконтентом. Кроме того, видеозаписи намного эффективнее, чем текст и статические изображения, для вызова таких эмоций, как страх, гнев или ненависть.

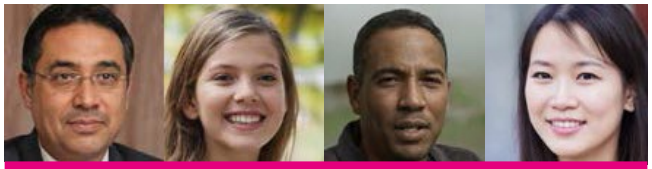
Тем не менее, также привлекли внимание некоторые примеры манипулирования изображениями/текстовым контентом на основе искусственного интеллекта. Что касается видео, основной целью алгоритмов обработки изображений является создание поддельного порнографического контента. Такие приложения, как DeepNude, могут за считанные секунды конвертировать фото в бикини в очень реалистичное изображение обнаженной натуры.

Неудивительно, что приложение работает только с изображениями женщин (любая попытка выбрать изображение мужчины сразу генерирует женские гениталии). Но это делает всех до единой женщин потенциальными жертвами «*порномести*», даже если никогда не существовало реальных обнаженных фотографий.

Изображение обнаженной натуры, созданное с помощью приложения deepnude.to

Эти нейронные сети не ограничиваются манипулированием изображениями реальных людей. Они также могут «*создавать*» совершенно новых людей или, по крайней мере, совершенно новые лица.

Коммерческое применение этой технологии очевидно: базы данных изображений можно рентабельнее заполнять с помощью ИИ, а не реальных людей. Но это также означает, что создавать поддельные профили в социальных сетях, например, с целью распространения политического контента, становится значительно проще.



Лица, сгенерированные случайным образом thispersondoesnotexist.com

Также были подозрения в попытках шпионажа с использованием изображений созданного компьютером профиля, например профиля LinkedIn некой «*Кэти Джонс*», предполагаемого исследователя, работающего в экспертно-аналитическом центре США.

Профиль в LinkedIn: Сгенерированный ИИ профиль для «*Кэти Джонс*»

До того, как экспертный анализ выявил несколько визуальных аномалий, предполагающих, что изображение было синтетическим, профиль успешно связывался с 52 политическими деятелями в Вашингтоне, включая заместителя помощника госсекретаря, старшего советника сенатора и известного экономиста.[4]

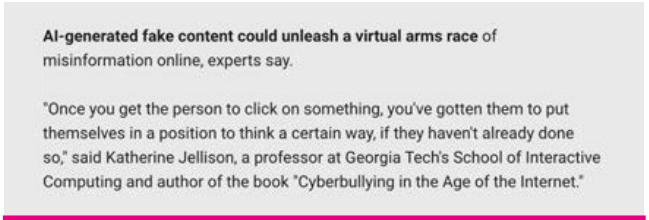
Аккаунт был быстро удален LinkedIn, но считается, что он принадлежал сети фантомных профилей, причем некоторые из которых все еще могут существовать, которые могли использоваться для фишинговых атак.

Тексты, сгенерированные ИИ

Описанное выше применение может быть особенно эффективно реализовано в сочетании с генерацией текста на основе ИИ.

Многие люди, возможно, уже слышали о такой возможности, благодаря текстовому генератору GPT-2, созданному исследовательской компанией OpenAI. Из-за возможности злоупотребления GPT-2 изначально считался слишком опасным, чтобы быть доступным для широкой общественности.[5] Позже компания решила опубликовать GPT-2 в несколько этапов, так как его создатели до тех пор не могли найти явных доказательств злоупотребления.[6]

Создатели признают, что даже если не было злоупотреблений, люди в значительной степени сочтут текст, сгенерированный GPT-2, достоверным, что генератор можно настроить для создания экстремистского контента и что идентификация сгенерированного текста будет сложной задачей. С помощью приложения «*Talk To Transformer*» любой желающий может попробовать GPT-2 для себя.



Выделенный текст введен, оставшийся текст генерируется ИИ с www.talktotransformer.de

При вводе одного или нескольких предложений в генератор им выводится фрагмент текста, начинающийся с введенных предложений. Результаты часто, но не всегда, на удивление последовательны. Они имеют ту же интонацию, что и введенные предложения, и имитируют достоверность за счет вымысла экспертов, статистических данных и цитат.

4 https://www.cnet.com/news/spy-reportedly-used-ai-generated-photo-to-connect-with-targets-on-linkedin/

5 https://openai.com/blog/better-language-models/

6 https://openai.com/blog/gpt-2-1-5b-release/

3.0 РАСПРОСТРАНЕНИЕ И ПОСЛЕДСТВИЯ

НАСКОЛЬКО ОПАСНЫ ДИПФЕЙКИ В РЕАЛЬНОСТИ?

3.1 Распространение

Точно оценить распространение дипфейков сложно, тем более, что их количество, несомненно, неуклонно растет.

Deeptrace, компания, предлагающая технологическое решение для обнаружения дипфейков, попыталась дать точную оценку в своем отчете: Состояние дипфейков: Ландшафт, угрозы и воздействие.[7]

По оценкам отчета, опубликованного в сентябре 2019 года, количество дипфейков почти удвоилось за семь месяцев с 7.964 в декабре 2018 года до 14.678 в июле 2019 года. Из этих дипфейков 96% было неконсенсуальным порнографическим контентом, который изображал исключительно женское тело.

Основными жертвами стали известные женщины, тысячи поддельных фотографий которых можно найти в Интернете. Согласно отчету Deeptrace, четыре самых популярных сайтов deep porn в одиночку зарегистрировали более 134 миллионов просмотров поддельных видеороликов со знаменитыми женщинами. Но многие частные лица также страдают от феномена порномести, упомянутого выше. Этот рост обусловлен в первую очередь большей доступностью инструментов и сервисов, которые позволяют создавать дипфейки без каких-либо знаний в области программирования.

В 2019 году также поступали сообщения о том, что для социальной инженерии используются сгенерированные ИИ языковые клоны. В августе газета The Wall Street Journal сообщила[8] о первом случае голосового мошенничества с использованием ИИ – также известного как «вишинг» (сокращение от «*голосовой фишинг*») – стоимостью 220.000 евро для немецкой компании, которая стала объектом атаки.

Программное обеспечение настолько успешно имитировало голос немецкого менеджера, в том числе его интонации и легкий немецкий акцент, что его британский коллега немедленно выполнил срочное требование звонящего о переводе указанной суммы. Хотя в настоящее время это единичный случай, похоже, что в будущем таких попыток будет больше.

Значительная часть охвата дипфейков в СМИ была сосредоточена на их способности дискредитировать политических оппонентов и подрывать демократические процессы. До сих пор этот потенциал не материализовался. Хотя были технически обработанные видеоролики с участием политиков, таких как Барак Обама, Дональд Трамп и Маттео Ренци, они в основном мотивировались сатирой или были созданы для демонстрационных целей, и их фальшь была быстро раскрыта.

---

7 Состояние дипфейков: Landscape, Threats, and Impact, Henry Ajder, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen, September 2019.

8 https://www.wsj.com/articles/fraudstersuse-ai-to-mimic-ceos-voice-in-unusualcybercrime-case-11567157402

«Хотя были технически обработанные видеоролики с участием политиков, таких как Барак Обама, Дональд Трамп и Маттео Ренци, они в основном мотивировались сатирой или были созданы для демонстрационных целей, и их фальшь была быстро раскрыта».

3.2 Последствия

Однако тот факт, что политики еще не использовали дипфейки для дезинформации, не означает, что дипфейки еще не повлияли на политический дискурс. Один пример, получивший мало внимания в западных СМИ, демонстрирует, как простое знание о существовании дипфейков может повлиять на политический климат.

Президент Габона Али Бонго не появлялся на публике в течение нескольких месяцев после перенесенного инсульта. Неудивительно, что начали распространяться слухи о том, что президент скончался. Чтобы опровергнуть домыслы, в декабре 2018 года президент опубликовал видеоролик, в котором произносил свою обычную новогоднюю речь. Но запись произвела противоположный эффект. Многие подумали, что Бонго выглядит странно, и сразу заподозрили, что видеоролик был подделкой. Вскоре после этого военные предприняли неудавшийся переворот, сославшись на предполагаемый дипфейк как на один из своих мотивов.[9]

Однако последующая судебно-медицинская экспертиза подтвердила подлинность записи. Али Бонго с тех пор оправился от инсульта и остается на своем посту.

Это показывает, что самая большая угроза, исходящая от дипфейков, заключается не в самих дипфейках. Сам факт того, что такие видеоролики технически возможны, поднимает вопрос: Можем ли мы по-прежнему доверять подлинности видео?

Этот вопрос бросит тень на президентские выборы в США 2020 года. В предвыборной кампании 2016 года уже начали играть роль дезинформация и манипуляции, поддерживаемые искусственным интеллектом, прежде всего в форме микроадресации и ботов. Дипфейки теперь представляют собой еще один инструмент в арсенале дезинформации. Даже если в избирательной кампании действительно используется мало дипфейков или вообще не используется, вполне вероятно, что многие политики с благодарностью примут возможность игнорировать настоящие, но неблагоприятные записи как подделки.

9 https://www.technologyreview.com/s/614526/the-biggest-threat-of-Deepfakes-isnt-the-Deepfakes-themselves/

3.3 Существуют ли какие-либо примеры положительного применения дипфейков?

«*Технология постоянно предоставляет нам способы причинить вред и поступить хорошо; это усиливает и то, и другое. [...] Но тот факт, что у нас также каждый раз появляется новый выбор, – это новое благо*»[10], – говорит Кевин Келли, давний главный редактор и один из основателей технологического журнала Wired. Может ли это утверждение относиться и к дипфейкам?

Технология особенно перспективна для киноиндустрии, в частности, для постпродакшена и дубляжа Почему? В настоящее время изменение фрагмента диалога задним числом очень дорого для киностудий. Необходимо перебронировать актеров, съемочную группу и съемочную площадку. Технология, лежащая в основе дипфейков, может позволить вносить такие типы изменений быстро и за небольшую часть стоимости.

Значительные улучшения могут быть внесены также в дублирование фильмов. Стало бы возможным адаптировать движения губ актеров к дублированным словам или синтезировать их голоса, чтобы адаптировать их к целевому языку, что означает, что дублирование больше не нужно.

Одним из примеров такого применения является видео с Дэвидом Бекхэмом, в котором он рекламирует кампанию против малярии.[11] Он «*говорит*» на нескольких языках, и в каждом случае его рот, кажется, идеально синхронизируется со словами.

10 **Цитата из** https://www.edge.org/conversation/kevin_kelly-the-technium/

11 https://www.malariamustdie.com/

Еще одна интересная область применения – образование: например, можно создавать видеоролики с историческими личностями, чтобы они рассказывали свои истории или отвечали на вопросы. Проект «*Dimensions of History*»[12] Фонда Шоа Университета Южной Калифорнии привлек большое внимание средств массовой информации, поскольку в нем были представлены интервью и голографические записи 15 выживших во время холокоста. Эта передвижная выставка экспонировалась в различных музеях США, а совсем недавно ее принимал Шведский исторический музей.

Посетителям выставки была предоставлена возможность задавать вопросы голограммам. Затем программа распознавания речи сопоставляла их вопросы с фрагментом интервью. С помощью дипфейковой технологии это можно реализовать в большем масштабе, на нескольких языках.

«Технология особенно перспективна для киноиндустрии, в частности, для постпродакшена и дубляжа».

12 https://sfi.usc.edu/dit

4.0 ПРОТИВОСТОЯНИЕ ДИПФЕЙКАМ

КАК МЫ МОЖЕМ ПРОТИВОСТОЯТЬ ПРОБЛЕМАМ, СВЯЗАННЫМ С ДИПФЕЙКАМИ?

Эти положительные примеры, конечно, не предназначены для сведения к минимуму потенциальных опасностей, связанных с дипфейками. Риски бесспорны и требуют решительных контрмер – по этому поводу существует консенсус. Но относительно точного характера этих контрмер нет единого мнения. Кроме того, возникает вопрос, как гарантировать права людей на свободу выражения мнения, не подрывая потребности общества в надежной информационной системе.

4.1 Технологические решения для выявления и борьбы с дипфейками

Один из подходов к борьбе с подделкой – разработка технологий, способных отличать поддельный контент от реального. Этот подход использует алгоритмы, аналогичные тем, которые изначально генерировали фейки. Используя GLTR, модель, основанную на упомянутой выше системе GPT-2, исследователи из MIT-IBM Watson AI Lab и HarvardNLP исследовали, можно ли использовать ту же технологию, которая используется для написания независимо сфабрикованных статей, для распознавания отрывков текста, сгенерированных ИИ. Когда в тестовом приложении создается текстовый отрывок, его слова выделяются зеленым, желтым, красным или пурпурным цветом, чтобы указать их предсказуемость в порядке убывания.

Чем больше доля слов с низкой предсказуемостью, а именно разделов, отмеченных красным и пурпурным цветом, тем выше вероятность того, что отрывок был написан автором-человеком. Чем более предсказуемы слова (и чем «*зеленее*» текст), тем выше вероятность, что текст был сгенерирован автоматически».

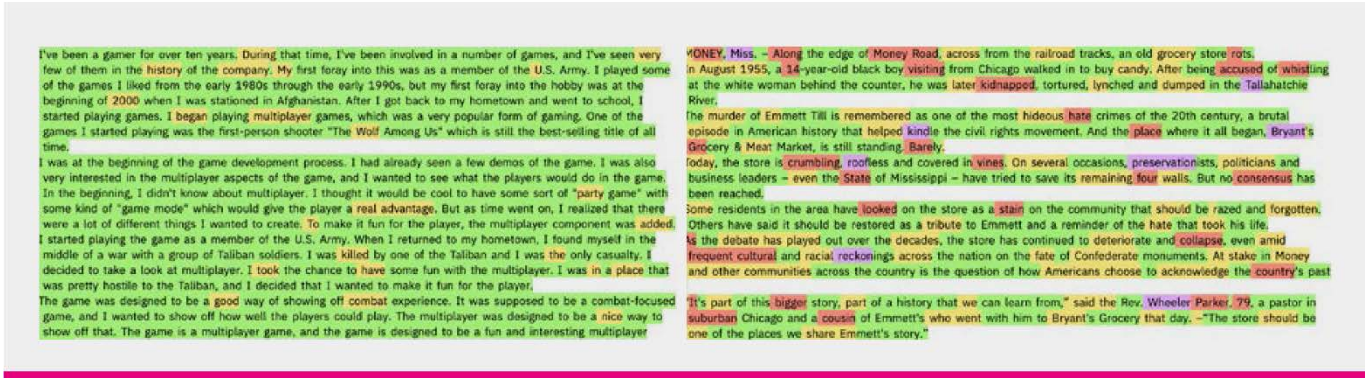
Подобные методы могут использоваться для выставления обработанных видеороликов. В 2018 году исследователи заметили, что актеры в дипфейковых видеороликах не моргали. Это было потому, что статические изображения, использованные для создания видео, в основном показывали людей с открытыми глазами. Но польза от этого наблюдения была недолгой. Как только эта информация стала достоянием общественности, стали появляться видеоролики с моргающими людьми. Подобную тенденцию можно ожидать и для любых других механизмов идентификации, обнаруженных в будущем. Эта игра в кошки-мышки ведется в области кибербезопасности на протяжении десятилетий, и прогресс всегда приносит пользу обеим сторонам.

Но это вовсе не означает, что следует прекратить попытки выявления дипфейков. В сентябре 2019 года Facebook в сотрудничестве с инициативой PAI[13], Microsoft и несколькими университетами объявил конкурс «*Deepfake Detection Challenge*»[14] с призовым фондом в 10 миллионов долларов.

---

13 Партнерство по ИИ (PAI) – это организация, объединяющая университеты, исследователей, НПО и предприятия для лучшего понимания влияния и воздействия ИИ на общество. www.partnershiponai.org

14 https://ai.facebook.com/blog/deepfake- detection-challenge/



Результаты анализа: автор-человек против текстового генератора, Источник: gltr.io

Facebook также ввел набор данных с изображениями и видео актеров, специально записанными для этой цели, чтобы в проекте были адекватные данные для работы. Несколько недель спустя Google также выпустил набор данных, содержащий 3.000 обработанных видео с той же целью.

Американское агентство по финансированию исследований DARPA также с 2016 года работает над распознаванием манипулируемого контента в рамках программы MediFor (сокращенно от Media Forensics), инвестировав более 68 миллионов долларов за два года.[15] Имеется мало информации о том, разрабатываются ли технические решения для борьбы с дипфейками в Германии и Европе, и если да, то какие именно.

Большинство мер предпринимается отдельными компаниями, такими как Deeptrace, которая была упомянута выше, а также исследовательскими проектами, такими как Face2Face Матиаса Ниснера[16], профессора Мюнхенского технического университета.

Согласно ответу правительства Германии на парламентский вопрос, представленный парламентской группой FDP, «Национальный исследовательский центр прикладной кибербезопасности» CRISP/ATHENE в настоящее время работает над этим вопросом с Мюнхенским техническим университетом и Институтом Фраунгофера.

Кроме того, немецкая международная вещательная компания Deutsche Welle (DW), Институт цифровых медиатехнологий Фраунгофера (IDMT) и Афинский технологический центр (ATC) запустили совместный исследовательский проект «*Digger*». Целью этого проекта является расширение веб-платформы проверки «*Truly Media*» от DW и ATC с помощью технологии криминалистической фоноскопии от Fraunhofer IDMT, среди прочего, для оказания помощи журналистам.[17] Однако этот ответ не предполагает какой-либо конкретной стратегии или намерений инвестировать в этот аспект со стороны федерального правительства.

«Чем больше доля слов с низкой предсказуемостью, а именно разделов, отмеченных красным и пурпурным цветом, тем выше вероятность того, что отрывок был написан автором-человеком. Чем более предсказуемы слова (и чем «зеленее» текст), тем выше вероятность, что текст был сгенерирован автоматически».

15 https://futurism.com/darpa-68-million-technology-Deepfakes

16 https://niessnerlab.org/projects/thies2016face.html

17 https://dip21.bundestag.de/dip21/btd/19/156/1915657.pdf

4.2 Попытки саморегулирования со стороны платформ социальных сетей

Хотя крупные технологические компании предоставили данные и финансовые ресурсы для технологического решения этой проблемы, призывы к Facebook и аналогичным компаниям принять дополнительные меры усиливаются, поскольку их платформы играют ключевую роль в распространении дезинформации. В ответ Twitter и Facebook опубликовали заявления о своих планах по устранению дипфейков в конце 2019 и начале 2020 года соответственно.

В ноябре 2019 года Twitter попросил своих пользователей высказаться по поводу «*предложения по политике в отношении синтетических и манипулируемых СМИ*». Затем в начале февраля 2020 года были объявлены руководящие принципы: любые фотографии, аудио- или видеоролики, которые были «*значительно изменены или фальсифицированы*» с целью ввести людей в заблуждение, будут удалены, если Twitter считает, что они могут причинить серьезный вред: например, поставив под угрозу физическую безопасность людей или провоцируя «*массовые гражданские беспорядки*». В противном случае твиты по-прежнему могут быть помечены как манипулируемые медиа, отображая предупреждение, когда контент пересылается, и снижая приоритет контента в пользовательских пересылках. Эти изменения вступают в силу 5 марта 2020 г.[18]

Is the media significantly and deceptively altered or fabricated?	Is the media shared in a deceptive manner?	Is the content likely to impact public safety or cause serious harm?	
✓	✗	✗	Content **may** be labeled
✓	✗	✓	Content is **likely** to be labeled, or **may** be removed.
✓	✓	✗	Content is **likely** to be labeled.
✓	✓	✓	Content is **very likely** to be removed.

:Twitter: новый подход к синтетическим и управляемым СМИ: https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html

18 https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html

«Имеется мало информации о том, разрабатываются ли технические решения для борьбы с дипфейками в Германии и Европе, и если да, то какие именно».

Facebook идет еще дальше. 6 января 2020 года Моника Бикерт, вице-президент Facebook по управлению глобальной политикой, объявила в своем блоге, что отныне дипфейки, отвечающие определенным критериям, будут удаляться с платформы.[19] Согласно сообщению в блоге, любой контент, измененный или синтезированный с помощью ИИ таким образом, что он кажется подлинным для обычного человека, будет удален. Однако сатирическое содержание исключено из этих руководящих принципов, что оставляет значительный простор для интерпретации.

Интересно, что рекомендации не относятся к чипфейкам; они явно и исключительно нацелены на контент, генерируемый ИИ. Соответственно, фальшивое видео Нэнси Пелоси, упомянутое ранее, по-прежнему доступно на Facebook.[20] Хотя Facebook признал, что его представители, проверяющие достоверность информации, отметили видео как фальшивое, он отказался удалять его, поскольку компания «*не применяет политику, требующую, чтобы информация, размещенная в Facebook, была правдивой*».[21]

Такой подход отражает позицию Facebook в отношении свободы выражения мнения и выходит за рамки вопроса о дипфейках. В ходе дискуссии о политической рекламе Роб Лезерн, директор по управлению продуктами в Facebook, написал в своем блоге в январе 2020 года, что такие решения не должны приниматься частными компаниями, «*поэтому мы выступаем за регулирование, применимое ко всей индустрии. В отсутствие регулирования Facebook и другие компании могут выбирать свою собственную политику*».

Безусловно, стоит обсудить, заслуживает ли внимания с этической точки зрения толкование Facebook свободы выражения мнения. Однако заявление Роба Лезерна обращает внимание на конкретный вопрос, а именно на отсутствие или, по крайней мере, неполноту регулирования.

---

19 https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/

20 YouTube, с другой стороны, еще одна платформа, которая способствует виральности ложной информации через свои рекомендательные алгоритмы, удалила видео, но отказалась сделать четкое заявление о том, как она будет обрабатывать дипфейки в будущем.

21 https://www.politico.com/story/2019/05/24/facebook-fake-pelosi-video-1472413

4.3 Попытки регулирования со стороны законодателей

В Германии дипфейки подпадают под «*общие и абстрактные правила*» в соответствии с ответом федерального правительства на краткий парламентский запрос, представленный парламентской группой FDP, как упоминалось выше. «*На федеральном уровне нет конкретных предписаний, которые охватывали бы исключительно дипфейк-приложения или были созданы для таких приложений. Федеральное правительство постоянно пересматривает правовую базу на федеральном уровне, чтобы определить, необходимы ли какие-либо корректировки для решения технологических или социальных проблем*».

Это означает, что некоторые частные аспекты вопроса, касающегося дипфейков, в том числе порномести, якобы косвенно охвачены существующими законами, но в действительности нет явного подхода к управлению манипулируемым контентом. Это относится ко всему спектру дезинформации в цифровом пространстве, а не только к частному случаю «*дипфейков*». Как отмечает автор исследования: «*Ответные меры регулирования на дезинформацию*»[22] из Stiftung Neue Verantwortung: «*предыдущие попытки регулирования и политических решений вряд ли подходят для сдерживания дезинформации*». Исследование юридической фирмы WilmerHale «*Законодательство в вопросе дипфейков: Общенациональный обзор*»[23] представляет подробный анализ статуса регулирования дипфейков в США.

В Соединенных Штатах явные части законодательства о дипфейках уже внесены в уголовное право: например, в Вирджинии, где неконсенсуальная дипфейковая порнография наказуема, и в Техасе, где любые дипфейки, предназначенные влиянию на избирателей, наказуемы. Аналогичное законодательство было также принято в Калифорнии в сентябре 2019 года.

Возможно, самое тщательное регулирование дипфейков было предпринято китайскими законодателями в конце 2019 года. Китайское законодательство требует, чтобы поставщики и пользователи онлайн-служб обмена видеосообщениями и аудиоинформацией четко отмечали весь контент, созданный или измененный с использованием новых технологий, таких как искусственный интеллект.

Хотя, безусловно, стоит подумать о том, могут ли аналогичные предписания также приниматься другими странами, случай с Китаем оставляет неприятное послевкусие: само правительство Китая использует дезинформацию на основе технологий, среди прочего, для целенаправленного вещания протестующим в Гонконге, и кажется неизбежным, что эти новые предписания будут использоваться в качестве предлога для дальнейшей цензуры.

Конечно, эффективно регулировать новые технологические явления непросто. В прошлом это часто оказывалось трудным. Например, чтобы водить автомобиль в Англии XIX века, в соответствии с Законом о локомотивах 1865 года перед автомобилем должен был идти второй человек, размахивая красным флагом.[24] Тем не менее, есть меры, которые законодатели уже могут предпринять, чтобы противодействовать явлению дипфейков. Так как 96% дипфейков в настоящее время относятся к неконсенсуальной порнографии, было бы хорошо начать явно наказывать за это, как это было сделано в Вирджинии и Калифорнии. Аналогичным образом можно регулировать вопросы, связанные с клеветой, мошенничеством и правами на конфиденциальность. Кроме того, законодатели должны разработать четкие руководящие принципы для цифровых платформ по единообразной обработке дипфейков в частности и дезинформации в целом.

Эти меры могут варьироваться от маркировки дипфейков как таковых и ограничения их распространения (исключения их из рекомендательных алгоритмов) до их удаления. Повышение медиаграмотности также должно стать приоритетом для всех граждан, независимо от возраста. Адекватное понимание того, как создаются и распространяются дипфейки, должно позволить гражданам распознавать дезинформацию и избегать введения в заблуждение.

22 https://www.stiftung-nv.de/sites/default/files/regulatorische_reaktionen_auf_desinformation.pdf

23 Matthew Ferraro, WilmerHale | Deepfake Legislation: A Nationwide Survey – State and Federal Lawmakers Consider Legislation to Regulate Manipulated Media.

24 https://sites.google.com/site/motormiscellany/motoring/law-and-the-motorist/locomotive-act-1865/

4.4 Ответственность личности: критическое мышление и медиаграмотность

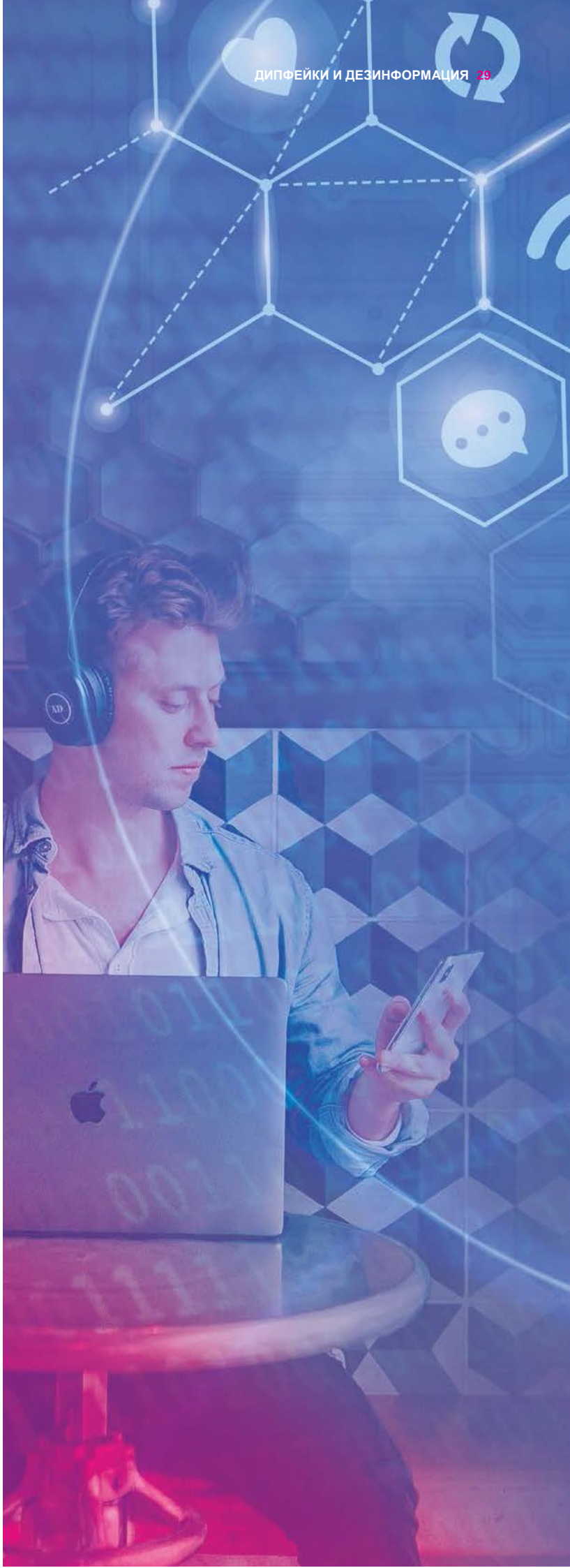
Критическое мышление и медиаграмотность – основа дифференцированного подхода к дезинформации. Конечно, невозможно и, вероятно, нежелательно просить каждого человека подвергать сомнению все, что он видит.

Но больше, чем когда-либо прежде, людям следует соблюдать осторожность при потреблении онлайн-контента. Самое простое, что может сделать каждый, если изображение, видео или текст кажутся подозрительными, – это задать поиск в Google. Часто это позволяет быстро снять маску с манипулируемого контента, поскольку детали манипуляции распространяются так же быстро, как и сам контент.

Это особенно важно для пользователей, которые хотят поделиться контентом, «*поставив лайк*» или комментируя его. Мы также можем уделять больше внимания тому, выглядят ли неестественно моргание, мимика или речь в видеоролике, размыты ли части изображения или не кажутся ли объекты не на своем месте.

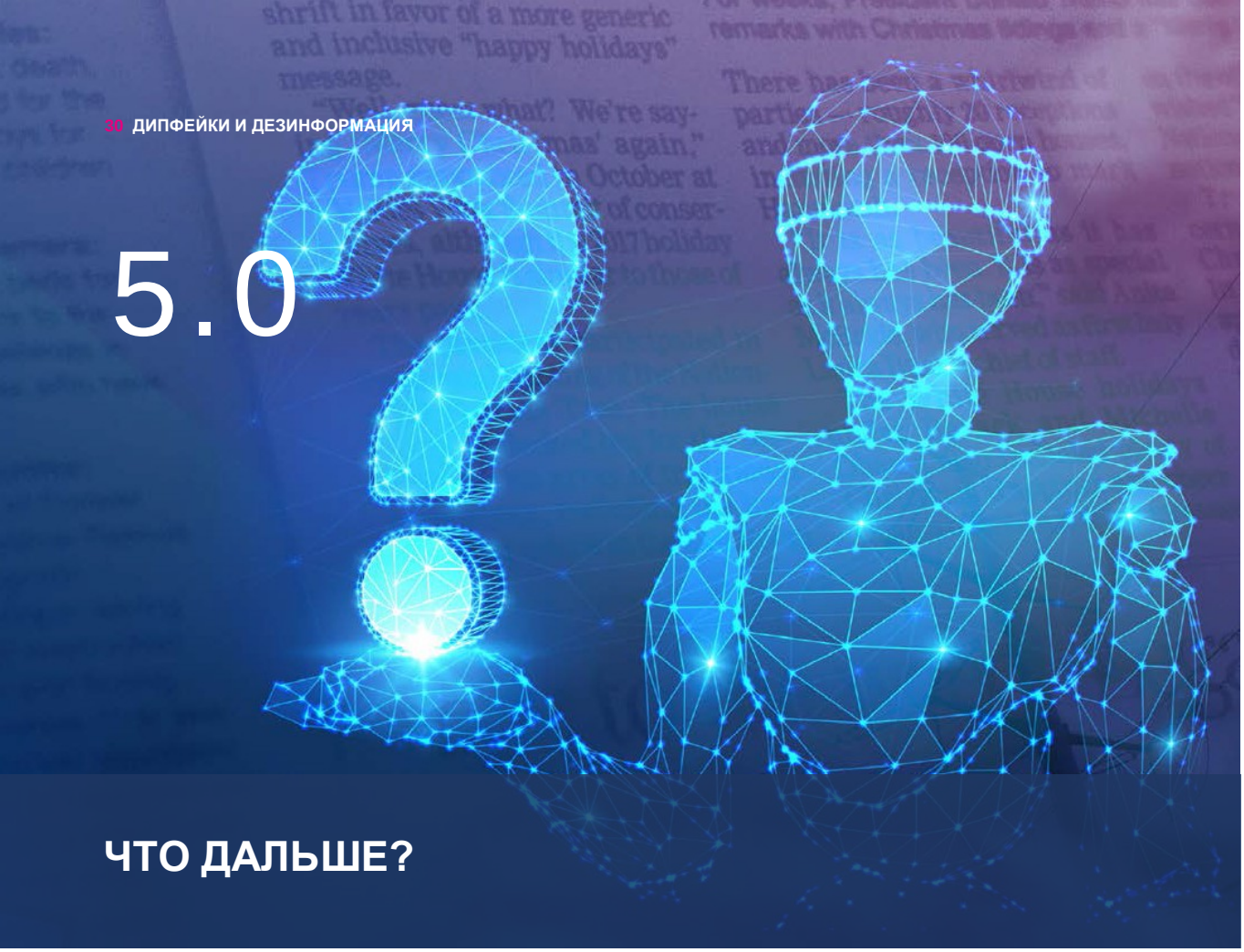
Однако эти подсказки быстро исчезнут по мере развития дипфейковой технологии. В будущем, вероятно, могут появиться надстройки браузера, которые будут автоматически идентифицировать манипулируемый контент и уведомлять пользователей подобно блокировщику рекламы. Но для этого в первую очередь необходимо осознавать возможность манипулирования контентом.

Для повышения осведомленности своих граждан Финляндия, страна, получившая наивысший рейтинг в исследовании устойчивости к дезинформации[25], предлагает возможности получения образования всему населению: от детсадовского до пенсионного возраста.



25 https://osis.bg/wp-content/uploads/2019/11/MediaLiteracyIndex2019_-ENG.pdf

5.0



ЧТО ДАЛЬШЕ?

Пока невозможно точно предсказать степень конкретного воздействия дипфейков на политику и общество, но это не оправдывает бездействие. Как подчеркивалось выше, ни фальшивые видео, ни дезинформация не являются новым явлением как таковым – новизна заключается в увеличивающейся простоте создания такого контента, его постоянно улучшающемся качестве и возможности его распространения.

Президентские выборы в США осенью 2020 года, несомненно, станут хорошей лакмусовой бумажкой. Тем не менее, рекомендация в данном случае не может быть: просто «*ждать и смотреть*».

Исследователи, технологические компании, журналисты, правительства и сами пользователи должны приложить все усилия, чтобы нейтрализовать негативное влияние поддельного контента. Первый шаг заключается во осуществлении явного регулирования и сильных контрмер против дипфейковой порнографии, так как это уже широко распространенное явление, которое наносит значительный вред его – в основном женщинам – жертвам.

Также требуются единые правовые нормы в отношении регулирования манипулируемым контентом в СМИ и на платформах социальных сетей. Мы не должны оставлять на усмотрение Facebook, Twitter, YouTube и других компаний решение, какой контент подпадает под действие свободы выражения мнения, а что выходит за ее рамки.

Эта задача является обязанностью законодателей и конституционной демократии. Однако не стоит поддаваться искушению полностью запретить дипфейки. Помимо рисков, технология открывает многообещающие новые возможности, в том числе в сфере образования, кино и сатиры. Сама технология нейтральна, это люди используют ее в пользу или во вред обществу.

Автор

Агнешка М. Валорска

Агнешка М. Валорска – эксперт по оцифровке и исполнительный директор консалтинговой компании в области управления и технологий Capco. Она руководила несколькими проектами по трансформации и инновациям для банков, страховых компаний, фармацевтических и автомобильных компаний.

Она основала консалтинговую компанию по цифровой стратегии CREATIVE CONSTRUCTION, которую Capco приобрела в 2020 году. Она особенно интересуется искусственным интеллектом и его влиянием на взаимодействие человека с машиной и, следовательно, на бизнес-модели и общество.

В рамках Digital Innovation Breakfast она организовала серию мероприятий с известными докладчиками по этим темам. Она опубликовала множество исследований и статей, регулярно выступает на конференциях и в компаниях.

Она написала главу книги об алгоритмическом обществе, в которой рассматривает этические вопросы искусственного интеллекта. Соответствующий том был опубликован научным издательством Springer в марте 2020 года.

Она изучала социальные и политические науки в Варшавском университете и Берлинском университете имени Гумбольдта, была стипендиатом Фонда Херти и Немецкого фонда академических стипендий.